Investigating Contextual Variations in Aviation Social Intention Recognition: A Scenario Design Framework for Evaluating Intelligent Systems

Omer Arslan

Supervisor(s): Hayley Hung, Vitaliy Popov, Arthur Mercier

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfillment of the Requirements
For the Bachelor of Computer Science and Engineering
January 25, 2026

Name of the student: Omer Arslan
Final project course: CSE3000 Research Project
Thesis committee: Hayley Hung, Ricardo Marroquim

Abstract

The design of data collection scenarios is critical for evaluating intelligent systems for social intention recognition in aviation. Identical aircraft behaviors can generate multiple equally plausible intention interpretations depending on situational context and the observer's professional perspective, yet existing research offers limited guidance for constructing scenarios that preserve this interpretive open-endedness. This study addresses this gap through an exploratory, literature-based investigation of how contextual factors shape intention interpretation across aviation roles. An integrated framework combining the 3Cs model of situational analysis and script theory is proposed to identify contextual dimensions influencing interpretation. Through qualitative synthesis of aviation literature, the framework demonstrates how variations in cues, classes, characteristics, and internal-external script configurations can produce divergent but valid intention narratives for the same observable behavior. The resulting scenario-first methodology provides structured guidance for designing aviation scenarios that support role-dependent intention annotation and evaluate intelligent systems. As a conceptual contribution, the framework requires empirical validation by aviation professionals.

1 Introduction

Intelligent systems operating in safety-critical aviation environments must interpret human social intentions under conditions of uncertainty, where misinterpretation can lead to severe consequences [1]. Crucially, the same observable aircraft behavior can give rise to multiple, equally plausible interpretations depending on contextual factors and the professional role of the observer [2]. For example, a rapid elevation change in an aircraft may be interpreted by air traffic control (ATC) as an emergency descent, by the flight crew as a deliberate weather-avoidance maneuver, or by maintenance personnel as an indicator of a potential system malfunction. None of these interpretations is inherently incorrect; rather, each reflects a valid perspective grounded in the observer's operational responsibilities and expectations.

This interpretive ambiguity presents a fundamental challenge for the design of intelligent systems in aviation. Traditional approaches to system evaluation often assume that observed aircraft behavior corresponds to a single underlying intention that can be treated as ground truth once subsequent actions are known [3]. Such assumptions overlook the fact that unrealized intentions (those that remain plausible but are never enacted) are an intrinsic aspect of real-world aviation operations [4]. As a result, intelligent systems trained using narrowly defined interpretations may perform adequately in expected situations, yet fail to reason effectively when faced with ambiguous behavior.

Although aviation research has long acknowledged that multiple stakeholders maintain distinct yet compatible understandings of the same operational situation, existing frameworks provide limited guidance on deliberately exposing this ambiguity during intelligent system design. Research on distributed situation awareness highlights the coexistence of role-specific perspectives within aviation teams, but focuses primarily on coordination and information sharing rather than systematic scenario construction [2; 5]. Similarly, models of pilot situation awareness demonstrate how interpretation shifts with context and workload, yet remain confined to a single professional role and do not address interpretive diversity across organizational boundaries [6]. Consequently, current literature lacks methodologies for designing data collection scenarios that deliberately capture multiple valid, role-dependent interpretations of identical aircraft behaviors.

In order to address this gap, the present paper investigates the following research question: **How can contextual variations in aviation be systematically characterized to guide AI researchers in designing data collection scenarios that capture multiple valid interpretations of social behavior?**

From the research question above, the following sub-questions are derived:

SQ1: What are the primary contextual dimensions identified in aviation literature that influence how aircraft behaviors are interpreted differently across aviation professional roles?

SQ2: How can external and internal scripts be used to structure and analyze divergent intention narratives in different aviation case studies?

SQ3: How can conflicts between internal professional scripts and external institutional scripts be systematically represented in aviation scenarios to reveal interpretive ambiguity?

SQ4: What are the limitations of using contextual framework and script-based scenario design to generate plausible intention explanations in aviation?

The primary contribution of this paper is an exploratory framework for designing scenarios that expose interpretive ambiguity prior to modeling. Specifically, this work proposes a scenario-first methodology that integrates the 3Cs framework (cues, characteristics, classes) with script theory to support the design of aviation data collection scenarios. The framework provides a structured approach to construct scenarios of graduated complexity, in which observable aircraft behaviors can be systematically interpreted through multiple professional perspectives. By varying contextual interpretations while holding observable cues constant, this methodology aims to support the development and evaluation of context-aware and explainable intelligent systems in aviation.

This paper proceeds as follows: Section 2 establishes the theoretical foundations. Section 3 presents the research methodology. Section 4 synthesizes insights from the aviation literature. Section 5 presents scenario construction through case studies illustrating variability in behavioral interpretation. Section 6 discusses the implications of these findings for intelligent system design and highlights gaps between current literature and scenario complexity. Section 7 addresses ethical considerations and reproducibility. Finally, Section 8 concludes the study and Section 9 outlines directions for future work.

2 Background Information

2.1 The 3Cs Framework

The 3Cs framework provides a systematic approach to analyzing how situations influence interpretation through three hierarchical levels of situational information [7]. This framework has been widely applied across various domains to understand how objective environmental features are transformed into subjective psychological meanings [7].

Cues represent the physical and objectively quantifiable stimuli accessible to all observers but carry no inherent psychological meaning [7]. In the scope of this paper, cues will be categorized into six key dimensions: persons, events, objects, activities, location, and time [7].

Characteristics capture the psychological meanings attributed to cues, encompassing dimensions such as threat, stress, task demands, and routine nature that vary across different observers [7]. They represent the critical transformation layer where raw sensory data gains psychological significance through individual interpretation processes.

Classes constitute the most abstract level, representing categorical judgments about entire situations that shape expectations about plausible actions and outcomes [7]. These high-level categorizations influence which behavioral repertoires are considered appropriate and which interpretation frameworks are likely to be activated.

2.2 Script Theory

Script theory provides a framework for understanding how individuals organize knowledge about how situations unfold and how actors should behave [8]. Scripts represent structured knowledge configurations that guide both understanding and action in familiar contexts [8].

Internal Scripts represent individual cognitive frameworks developed through training and experience, operating through hierarchical components: play components (overall knowledge about the collaborative performance), scene components (knowledge about specific situations within broader contexts), role components (understanding of participant roles and their activities), and scriptlet components (knowledge about specific activity sequences within scenes) [8].

External Scripts represent institutionalized frameworks codified in regulations and procedures, establishing expectations through corresponding scaffolds: play scaffolds (general task definitions), scene scaffolds (structured sequences of phases), role scaffolds (assigned responsibilities), and scriptlet scaffolds (specific prompts for activity sequences) [8].

The interaction between internal and external scripts creates a dynamic tension that generates interpretive variation. When internal scripts align with external scripts, interpretation tends toward consensus [8]. However, when these frameworks diverge, multiple valid interpretations emerge. This interpretive variability represents what can be termed **open-endedness**: the breadth of possible ways behavior can be explained by different intentions [8]. This phenomenon, which arises from conflicting internal scripts, mismatches between internal and external scripts, or simultaneous intention processes, is the key focus that this paper's methodology aims to systematically capture and analyze.

2.3 Multi-Perspective Annotation and Data Collection

In machine learning, annotators provide labels or interpretations of data for training intelligent systems. Traditional annotation approaches aim to minimize inter-annotator disagreement through majority voting to establish single, ground-truth labels [9]. However, annotator characteristics create systematic and legitimate differences in how identical scenarios are interpreted. These characteristics include professional roles, training backgrounds, and experiences. The open-endedness of intention interpretation comes directly from the annotators themselves rather than treating disagreement as noise to be eliminated [9].

Multi-perspective annotation deliberately samples annotators from different professional groups to capture legitimate interpretive variation [10]. In aviation, this means sampling from pilots, ATC, maintenance personnel, and other relevant stakeholders. Each group brings distinct internal script developed through their specialized training and operational experience. When presented with identical environmental cues, these groups generate different psychological characteristics interpretations, leading to different situation classifications [9]. This systematic variation in the 3Cs framework enables the manipulation of interpretive open-endedness. A scenario annotated only by ATCs produces closed-ended interpretations reflecting a single professional script. The same scenario annotated by both controllers and pilots becomes open-ended, capturing multiple valid intention narratives based on their different script configurations. This approach enables intelligent systems to be trained on datasets that preserve rather than collapse the collaborative interpretive processes essential for aviation safety operations.

3 Methodology

This research adopts an exploratory, discovery-oriented approach to examine contextual variations in the interpretation of social intentions in aviation. Given the absence of established frameworks for systematically designing scenarios, a qualitative literature synthesis is used as the primary methodological instrument. The insights derived from this survey are subsequently used to conduct case study development.

3.1 Literature Survey Design

The literature survey is structured in two complementary parts with distinct inclusion criteria. The first phase identifies literature providing conceptual leverage on intention interpretation, contextual variability, and script-based reasoning. Included papers describe real or hypothetical situations involving observable behavior that can plausibly be interpreted as intentional. While social situations are prioritized, individual-focused studies are retained when intelligent systems are involved or when the described behavior plausibly scales to a social aviation context. These studies received less analytical emphasis in the survey and were primarily used to anchor the discussion to the current state of the art.

The second part of the survey focuses on literature that directly supports scenario construction for data collection. Included papers describe short-term (5-30 minute), real or hy-

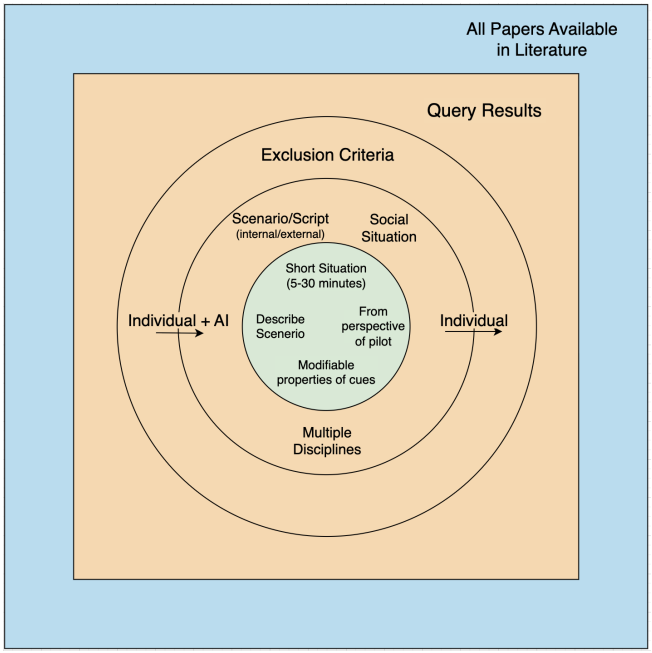

Figure 1: Literature filtering process for aviation social intention research. The concentric circles illustrate progressive filtering from complete literature (outer ring) to final inclusion criteria (center). Papers must describe social situations with internal/external scripts, involve short-term scenarios from pilot perspectives, and contain modifiable situational properties for systematic scenario variation.

Subjectivity Annotation	**Behavior Understanding**	**Domain Specific Terms**	**Exclusion Criteria**
social intention*	subjective annotation	aviation	drone*
intent* recognition	plausible narrative*	aircraft	uav
intent* estimat*	alternative narrative*	flight deck	unmanned
intent detection	script theory	cockpit	optimization
intent* predict*	social norm*	pilot*	
behavio*r interpret*	contextual variation	flight simulation	
behavio*r recogn*	human factors	aviation simulation	
social percept*	subjectivity	flight scenario*	
action prediction	perspectiv*		
goal inference	multi-perspective		
narrative expl*	ambigu*		
plausible narrative*	uncertainty		
intent* inferenc*	context model*		
goal recogn*	situation awareness		
goal estimat*	external script*		
goal inferenc*	internal script*		
behavio*r estimat*	situational script*		
behavio*r inferenc*	situation*		
	case stud*		
	scenario*		
	multiple annotator		
	perspective*		

Table 1: Search Query Clusters Used in the Literature Review

pothetical social situations in which an intelligent system participates as a passive observer or an active intervening agent. Intention interpretation must be analyzable from the pilot's perspective within the scene. Across both parts of the survey, attention is given to identifying modifiable properties of situational cues, mainly scene structure, actor roles, and contextual signals, that can be systematically varied to generate alternative, yet plausible, intention interpretations.

3.2 Search Strategy and Query Construction

To execute the approach described in *subsection 3.1*, a structured search was conducted using Scopus, following the multi-step filtering process illustrated in *Figure 1*. The search strategy progressively narrowed an initially broad conceptual space into a focused corpus of aviation-relevant studies addressing intention interpretation under contextual uncertainty.

The search query was designed to be systematic and reusable across multiple application domains within the broader research program, including aviation, restaurant service, manufacturing, hospital environments, and driving. Its core conceptual components were standardized across domains, with domain-specific constraints applied as needed, enabling methodological comparability across domains while allowing the present study to focus exclusively on aviation.

As shown in *Table 1*, the query consisted of four conceptual clusters. The first cluster targeted intention and behavior-related concepts associated with social intention, intention recognition, and behavior interpretation, capturing literature on how observers infer latent intentions from observable actions. The second cluster addressed domain-general contextual and interpretive factors, such as subjectivity, ambiguity, script theory, and situation awareness, drawing from interdisciplinary traditions, such as psychology, anthropology.

The third cluster constrained the search to operational aviation contexts using domain specific terms, such as aircraft, cockpit, and flight simulation, ensuring relevance to manned aviation operations. These terms were optimized to balance feasibility and coverage, producing a set of papers small enough for manual inspection of abstracts and introductions, yet large enough to avoid information loss.

The fourth cluster applied exclusion criteria to remove studies outside manned aviation and human-centered intention interpretation. The terms **drone** and **UAV** were excluded because such research typically focuses on autonomous systems, algorithmic planning, or navigation without a human-decision maker, limiting relevance to cockpit human–machine interaction. The term **unmanned** was excluded to prevent including hybrid system studies that nonetheless lack a human pilot as the primary intentional agent within the operational loop. Finally, the term **optimization** was excluded to remove studies focused on mathematical control that ignore subjective intention attribution.

The search query shown in *Table 1* resulted in an initial set of 165 papers from Scopus. These papers were subjected to a three-stage screening process consistent with the filtering structure illustrated in *Figure 1*.

In the first screening stage, abstracts and titles were manually reviewed to assess whether each paper described aviation contexts involving human decision-makers interacting with intelligent systems. Papers focused on unmanned systems, mathematical optimization without human factors, or non-aviation domains were excluded, reducing the corpus to 129 papers. At this stage, studies describing individual sit-

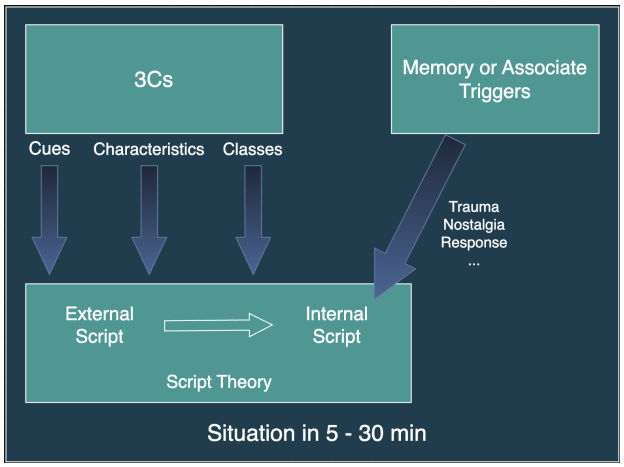

Figure 2: Integrated 3Cs and script theory framework for aviation scenario analysis. The diagram illustrates how environmental situation modeling combines with cognitive interpretation processes (internal and external scripts) to systematically identify sources of interpretive variation.

uations were retained if they plausibly informed intelligent system behavior, serving to anchor the state-of-the-art.

The second screening stage involved examining abstracts and introductions to identify papers describing concrete social situations where intelligent systems acted as passive observers or active participants alongside human aviation professionals. Studies addressing extended mission-level analysis rather than operational episodes of 5-30 minutes were excluded. This stage reduced the corpus to 78 papers.

In the final screening stage, full papers were examined to select those containing explicit aviation scenarios where: (1) intelligent systems intervened live or perceived and intervened post-hoc, (2) behavioral cues could be observed by multiple parties including human professionals and intelligent systems, and (3) intention interpretation could reasonably be analyzed from the human operator's perspective. Papers meeting these criteria formed a final corpus of 58 human–AI aviation interaction studies for detailed analysis.

3.3 Contextual Dimensions Framework

This research contributes a methodology that integrates the 3Cs framework with script theory to systematically categorize sources of interpretive variation in aviation scenarios. While environmental situational modeling captures objective contextual factors, it does not account for the cognitive dynamics through which aviation professionals interpret identical behavioral cues using distinct internal and external scripts. By integrating environmental cues, characteristics, and classes with role-specific script configurations, the framework provides a structured means of revealing multiple valid interpretations for identical observable behaviors (see *Figure 2*).

The framework supports systematic scenario construction by structurally transforming literature insights into data collection scenarios. Aviation behaviors observable to multiple professional roles are identified from the literature survey, with emphasis on behaviors that plausibly support divergent interpretations across stakeholders. Each behavior is analyzed using both 3Cs and script theory components to catalog environmental cues and examine how their psychological characteristics and script configurations are mapped across different aviation roles to identify points of alignment and conflict. Such misalignments are treated as primary sources of interpretive ambiguity.

Scenarios are operationalized by holding environmental cues constant while systematically varying characteristic interpretations and script configurations across roles. This manipulation allows the same observable behavior to support multiple plausible intention narratives, reflecting the open-endedness inherent in real-world aviation operations. Each scenario is temporally scoped to match the typical operational episodes of 5–30 minute. This way, intentions remain measurable within the scenario window.

4 Literature Survey

This literature synthesis examines the 58 papers identified through the structured search process to reveal a central finding: in aviation, intention recognition is not a direct decoding of behavior but a complex interpretive process mediated by professional scripts. Studies focused on **Non-Intelligent Systems** is synthesized in *subsection 4.1* to establish the mechanisms of script-mediated interpretation. Studies on **Intelligent Systems** are then critiqued in *subsection 4.2* for their systemic failure to model this essential human process.

4.1 Non-Intelligent Systems

The core of interpretive ambiguity in aviation stems from how human professionals imbue objective environmental data with subjective meaning. The following sections break down the mechanisms driving this process, from foundational cognitive acts to complex conflicts between institutional and individual worldviews.

The Foundation: From Objective Cues to Subjective Characteristics

The most fundamental source of interpretive variation occurs when observers apply different internal scripts to the same objective cue, leading to systematically opposing psychological interpretations [11]. Research on fighter pilots demonstrates that extended instrument scanning patterns of 2-3 seconds followed by rapid control input sequences constitute identical, observable cues that can be interpreted through opposing professional frameworks [12]. However, the meaning attributed to these cue diverged based on the observer's internal script. An experienced pilot internal script interpreted the scanning-input sequence as evidence of high situational awareness and decisive action under pressure, reflecting comprehensive situation assessment followed by confident execution. Conversely, an overwhelmed pilot internal script interpreted the identical behavioral sequence as task saturation and reactive response patterns, indicating extended information seeking followed by rushed action due to cognitive overload [12].

This finding demonstrates that the characteristics layer represents the critical transformation point essential for human-level understanding of aviation professional behavior. Studies of eye tracking in military aviation environments reveal that

ocular parameters like fixation rates correlate significantly with flight parameters during different operational phases, yet these correlations vary systematically based on pilot experience level and training background [12; 13; 14].

Cross-Role Conflicts: When Professional Scripts Collide
Interpretive ambiguity magnifies in multi-role aviation environments where professionals operate with conflicting scripts. Crew Resource Management research demonstrates that aviation safety culture has evolved through six distinct historical training generations, each emphasizing different script priorities: first-generation scripts focused on individual authority, while sixth-generation scripts prioritize systemic threat management and distributed decision-making [15]. These generational differences create predictable script conflicts when aviation professionals from different training eras interpret identical operational behaviors.

The forensic analysis of US Airways Flight 1549 provides empirical evidence of legitimate cross-role script conflicts [16]. During the emergency, Captain Sullenberger's 8-second delay in responding to an ATC frequency change request constituted an objective cue observable to all parties, yet generated two valid but opposing interpretations:

- ATC personnel, applying traffic management scripts developed through training that prioritizes immediate communication for orderly traffic flow, interpreted the delay as non-compliance and reduced situational awareness [17]. Their internal scripts, optimized for system efficiency, could not accommodate the possibility that aircraft control might legitimately take precedence over communication protocols during emergency situations.
- Flight crew members, applying aircraft emergency scripts dictating aircraft control takes precedence over non-critical communications, interpreted the identical delay as appropriate prioritization and high workload management [17]. Their internal scripts, developed through emergency training scenarios and operational experience, recognized the delay as evidence of proper task prioritization rather than performance degradation.

This case exemplifies how the classes dimension operates: the same event receives different categorical classifications ("routine traffic management" vs. "aircraft emergency"), leading to opposing but professionally legitimate judgments. Research on naturalistic decision making in aviation demonstrates that experienced pilots use Recognition-Primed Decision Making strategies involving situation recognition, serial option evaluation, and mental simulation rather than analytical comparison of alternatives [18; 19]. These cognitive strategies create legitimate interpretive variation when different professional roles assess identical pilot behaviors through their respective decision-making frameworks.

External Script Conflicts: The Influence of Training and Institutional Norms
Catastrophic interpretive failures arise from conflicts between individuals' internalized scripts and formalized external scripts. Aviation research identifies "training syndrome" as a documented source of misinterpretation where behaviors learned in simulated contexts become inappropriately applied to operational environments [13; 15]. This phenomenon occurs when training-based external scripts create internal script configurations that prove counterproductive during actual operational scenarios [15].

The Air France 447 accident provides a paradigmatic example of external script conflict. The pilot's continuous nose-up inputs during an aerodynamic stall, which appear incomprehensible from an operational flight safety perspective, become interpretable through training script analysis [20; 21]. The pilot's internal script, shaped by simulator-based emergency procedures emphasizing altitude maintenance, conflicted with operational requirements for nose-down stall recovery. Limited high-altitude manual flying experience caused training scripts to dominate operational decision-making during the critical period, creating a script configuration mismatch that contributed to the fatal outcome.

Similarly, the Tenerife disaster demonstrates external script conflicts in communication-critical environments [10]. The KLM captain, with extensive flight instructor experience, operated according to training environment external scripts where communication is compressed and controller statements like "Okay" indicate implicit approval [10]. When the ATC transmitted "Okay-stand by for takeoff," script-driven expectation bias led to selective attention focusing on "Okay" confirmation while filtering the "stand by" directive. Operational ATC protocols, however, require external scripts emphasizing clearance separation and explicit authorization requirements. Standard ATC communication external scripts dictate that "stand by" represents an explicit and overriding command to wait, regardless of preceding acknowledgment language. The conflict between compressed training external scripts and redundant operational external scripts created interpretive ambiguity that contributed directly to the collision.

4.2 Intelligent Systems

This subsection examines documented aviation contexts where pilots interact with intelligent systems, analyzing how these systems fail to account for the interpretive complexity that aviation professionals consider essential for comprehensive situational awareness.

Outcome-Based Training and Script Blindness
Contemporary aviation AI systems rely on outcome-based training methodologies that associate observable pilot behaviors with post-hoc outcomes rather than capturing the real-time intention formation processes central to professional aviation practice [22; 23]. This approach creates systematic script blindness—the architectural inability to model how different professional training scripts generate legitimate interpretive variation from identical behavioral cues [24].

At the cue level, outcome-based systems correctly observe objective pilot behaviors such as instrument scanning sequences, and procedural timing patterns. However, these systems systematically bypass the characteristics layer where aviation professionals assign psychological meanings like workload management, threat assessment, or decision confidence to identical behavioral patterns [25]. Instead, systems proceed directly from cues to situation classes based solely on eventual outcomes, eliminating the interpretive richness that

allows multiple valid intention narratives to coexist.

Script blindness manifests most clearly in systems' inability to handle unrealized intentions: valid intentions that remain plausible throughout an observational episode but are never enacted due to changing circumstances [25]. Consider a pilot monitoring unstable approach indicators who maintains go-around intention readiness while continuing the approach. At the cue level, this produces observable behaviors: extended instrument scanning, reduced communication, and delayed configuration changes. At the characteristics layer, experienced pilots would interpret these cues as evidence of heightened situational awareness and appropriate caution, while training-oriented observers might see hesitation or indecision. However, if conditions improve and landing is completed, outcome-based training forces a single situation class of "landing intention," systematically eliminating the coexisting "go-around readiness" narrative that experienced aviation professionals consider essential for comprehensive situational awareness.

Empirical Evidence of Script Blindness

Mao et al.'s implementation of a pilot intent recognition system provides concrete evidence of script blindness. Their system monitored pilot operational sequences across five aviation tasks (threat identification, weather avoidance, route optimization, instrument landing, and communication coordination) using 46 cockpit control inputs, achieving 88.89% accuracy through operation matching, sequence matching, and coverage rate metrics [24]. Despite this performance, systematic analysis reveals fundamental limitations. At the cue level, the system correctly observes shared, objective stimuli: specific button press sequences, timing patterns between inputs, and procedural execution orders that are accessible to both human observers and the AI system. Paper mentions that a sequence of autopilot disconnect, manual control inputs, and communication button activation constitutes identical cues observable by all parties [24].

The critical failure occurred at the characteristics layer, where the system treated cues uniformly regardless of the psychological meanings that different internal scripts would assign [26]. In emergencies, the identical button press sequence represented fundamentally different characteristics:

- Experienced Emergency Script: Decisive expert action reflecting comprehensive situation assessment and confident execution under pressure
- Overwhelmed Reactive Script: Task saturation response indicating rushed information seeking followed by reactive communication due to cognitive overload
- Training-Based Script: Procedural compliance demonstrating systematic adherence to memorized emergency sequences regardless of situational appropriateness

The system's uniform processing eliminates these interpretive distinctions, proceeding directly to situation classes like "emergency response" or "threat management" without modeling the script-mediated reasoning that aviation professionals consider essential for appropriate response coordination.

The system also fails to model conflicts between external scripts embedded in the AI's training data and pilots' internalized scripts developed through diverse operational experience [24]. The system's external script prioritizes operation matching and sequence coverage, potentially conflicting with pilots' internal scripts that emphasize situational adaptation and context-sensitive decision making over procedural consistency.

These failures indicate that evaluation scenarios must vary pilot experience levels, training backgrounds, and operational contexts while holding control input sequences constant. Scenarios should present identical cues that support multiple valid characteristics interpretations to test whether systems can distinguish between these fundamentally different intention formation processes.

Cockpit AI Assistance Systems

The HAIKU project's UC1 (cockpit warning management) and UC2 (weather data sharing) systems exemplify failures at the characteristics layer, where identical cues receive uniform interpretation regardless of pilots' internal scripts [23]. However, those systems bypass the interpretive process where human pilots, guided by their internal scripts, would assign different psychological characteristics to these cues.

At the cue level, UC1 correctly processes shared, objective stimuli: instrument warnings accessible to both pilots and the AI system. However, the system's failure occurs at the characteristics layer, where identical cues should receive different psychological interpretations based on pilots' internal scripts:

- Experienced Emergency: Interprets instrument warnings as confirmatory information for situation assessment [23].
- Overwhelmed Startle: Interprets identical warnings as competing attentional demands requiring immediate sequential processing [23].
- Training-Based Procedural: Interprets warnings as checklist triggers demanding systematic scanning regardless of flight phase criticality [23].

Findings of the project demonstrated that UC1's external script of "optimal scanning prompts" has conflicted systematically with the internal scripts above, forcing a single situation class of "systematic warning response" that eliminates the legitimate interpretive variation [23].

At the cue level, UC2 processes objective weather data. The characteristics layer failure manifests when the system's optimization logic conflicts with pilots' internal scripts:

- Safety-First: Interprets weather data as threat indicators requiring maximum operational margins
- Operational Efficiency: Interprets identical data as tactical information for fuel and schedule optimization
- Experience-Based Contextual: Interprets data through accumulated knowledge of specific aircraft performance and regional weather patterns

UC2's external script enforces algorithmic optimization, proceeding directly to situation classes like "weather avoidance" or "route efficiency" without modeling the script-mediated reasoning processes that pilots use to balance competing operational priorities.

Observer	Internal Scripts	External Scripts	Characteristics	Classes
Experienced ATC Routine Context	Prioritizes traffic flow, predictability, and immediate compliance for system efficiency.	Institutional procedures requiring standardized phraseology and timely pilot responses.	**Non-compliance.** A minor, but notable, deviation from the norm.	Routine Traffic Violation
Experienced ATC Emergency Context	Prioritizes threat detection and airspace safety; non-compliance is a potential indicator of a critical event.	Emergency protocols that elevate the significance of any deviation from expected behavior.	**Potential Incapacitation.** A critical indicator of a possible loss of crew situational awareness.	Undeclared Emergency
Experienced Pilot High-Stress Context	Prioritizes the "Aviate, Navigate, Communicate" hierarchy; aircraft control is paramount.	Emergency checklists and company policies that mandate stabilizing the aircraft before external communication.	**Appropriate Task Prioritization.** A deliberate and disciplined action.	Effective Workload Management
Novice Pilot High-Stress Context	Tends toward rigid procedural adherence; may struggle to prioritize under high cognitive load.	Training-based procedures that may not have been fully internalized for dynamic, high-stress situations.	**Hesitation/Indecision.** A sign of being overwhelmed or task-saturated.	Potential Task Saturation

Table 2: Multi-Dimensional Script-Based Interpretations of the 8-Second Delay

5 Scenario Construction

Following the methodology established in Section 3, this section develops the scenarios that systematically vary classes, characteristics and script configurations while holding environmental cues constant, creating the interpretive open-endedness essential for evaluating intelligent systems' ability to preserve multiple plausible intention narratives.

5.1 Case Study A: The Communication Delay

This scenario models the cross-role script conflicts common in aviation, inspired by the forensic analysis of US Airways Flight 1549 [16]. It is structured to generate multiple levels of open-endedness by varying not only the observer's role but also their experience and the perceived emergency context.

Observable Cue Identification

- **Cue:** During flight operating at 35,000 feet in moderate traffic density, ATC issues a routine frequency change instruction. The flight crew delays acknowledgment for 8 seconds while maintaining stable flight parameters.

Multi-Dimensional Variation and Script Analysis

Two primary observer roles are involved whose perspectives could be sampled for data collection annotation: the ATC and the two pilots controlling the plane. Table 2 maps the cue to divergent interpretations by explicitly detailing the internal and external scripts that guide each observer's reasoning process.

Intention Narratives Formulation for Evaluation

- **Narrative 1 (Experienced ATC, Routine):** "The pilot has momentarily failed to comply, likely due to a minor distraction. The intention is to respond shortly, but this represents a minor lapse in protocol."
- **Narrative 2 (Experienced ATC, Emergency):** "The pilot's failure to respond suggests a critical event in the cockpit. Their intention is unknown, and they may be unable to communicate, making the aircraft an unpredictable risk."
- **Narrative 3 (Experienced Pilot, High-Stress):** "The pilot is intentionally and correctly prioritizing aircraft control over a non-critical communication. The intention is to ensure flight safety first, which is the correct procedure."
- **Narrative 4 (Novice Pilot, High-Stress):** "The pilot may be struggling to manage multiple tasks. The intention to communicate is likely present but delayed due to cognitive overload, increasing the risk of error."

5.2 Case Study B: The Stall Recovery Input

This scenario models the deep ambiguity that arises when a pilot's actions could be attributed to profound human error or a rational response to faulty system information, inspired by incidents like Air France Flight 447 [20; 21].

Observable Cue Identification

- **Cue:** During a high-altitude aerodynamic stall warning, the pilot applies continuous nose-up stick input.

Multi-Dimensional Variation and Script Analysis

Table 3 details how different professional roles and assumptions on system reliability lead to opposing conclusions, explicitly tracing the logic back to the guiding internal and external scripts.

Intention Narratives Formulation for Evaluation

- **Narrative 1 (Instructor, Reliable Sensors):** "The pilot is panicking. Their intention is irrational and directly contrary to their training, making recovery impossible."
- **Narrative 2 (Instructor, Faulty Sensors):** "The pilot likely intends to stop a perceived rapid descent shown by faulty instruments. Their intention is to save the aircraft, but their actions are based on dangerously incorrect information."
- **Narrative 3 (Engineer, Reliable Sensors):** "The system is performing as designed. The pilot's intention is unclear and their actions are the source of the failure."
- **Narrative 4 (Engineer, Faulty Sensors):** "The pilot is likely intending to climb to a safe altitude, acting rationally based on the data presented to them. The root cause is a system malfunction, not pilot error."

Observer	Internal Script	External Script	Characteristics	Classes
Pilot / Instructor (Assuming Reliable Sensors)	Built on the ingrained "push, roll, power" mantra for stall recovery; prioritizes immediate procedural execution.	Standard Operating Procedures and flight manuals that explicitly mandate a nose-down input to recover from a stall.	**Gross Pilot Error.** A catastrophic failure to apply fundamental training.	Loss of Control by Pilot Error
Pilot / Instructor (Suspecting Faulty Sensors)	Experience-based knowledge that high-altitude sensor errors can produce misleading warnings and data.	Specialized training on system limitations and unreliable airspeed checklists that override basic stall recovery mantras.	**Misguided Procedural Adherence.** A rational attempt to fly the aircraft based on faulty instrument readings.	System Induced Pilot Error
Maintenance Engineer (Assuming Reliable Sensors)	System-first diagnostic mindset; assumes pilot competence and searches for external causes or anomalous human factors.	Maintenance logs and system performance data showing no history of sensor faults for this specific aircraft.	**Anomalous Pilot Action.** A puzzling behavior that contradicts expected competency.	Unexplained Pilot Deviation
Maintenance Engineer (Suspecting Faulty Sensors)	Diagnostic mindset that prioritizes potential hardware/software failures as root causes for pilot actions.	Diagnostic fault trees and airworthiness directives that provide knowledge of potential sensor failures.	**Logical Response to Faulty Data.** A rational action given the high probability of erroneous sensor data.	Suspected Sensor Failure

Table 3: Multi-Dimensional Script-Based Interpretations of the Nose-Up Input

6 Discussion

The systematic analysis of the two case studies demonstrates that open-endedness in aviation intention interpretation is a manipulable, and not binary, characteristic. Both cases reveal predictable patterns across scenario parameters that have direct implications for intelligent system design.

Case Study A highlights an inverted U-shaped relationship between temporal ambiguity and interpretive diversity. The 8-second communication delay represents a zone of optimal ambiguity, long enough to violate routine operational expectations but insufficient to trigger universal emergency protocols. Within this window, multiple valid interpretations coexist: experienced pilots may invoke workload management scripts, while ATCs apply compliance monitoring scripts, generating legitimate cross-role tension. Shorter delays (2–3 seconds) collapse interpretations toward routine operations consensus, whereas longer delays (15+ seconds) force convergence on emergency protocols. This illustrates how the characteristics layer in the 3Cs framework transforms identical cues into divergent psychological meanings, which are further modulated by observers' internal and external scripts.

Case Study B reveals that open-endedness scales primarily with uncertainty about system reliability. When observers assume fully reliable sensors, a pilot's nose-up input during a stall warning yields minimal interpretive variation: professional scripts converge on procedural violation. Introducing moderate uncertainty generates multiple plausible narratives, ranging from rational response to system-induced error. This demonstrates that interpretive diversity arises not solely from observable behavior, but from observer confidence.

A cross-scenario analysis identifies a constraint hierarchy governing interpretive outcomes. Strong constraints override role- and experience-based differences, producing closed-ended interpretations. Weak constraints allow scripts to dominate, generating high open-endedness. Moderate constraints, however, create competing interpretations that cannot be resolved through additional information. Historical aviation incidents illustrate this principle: in US Airways Flight 1549, an 8-second communication delay produced conflicting but professionally legitimate interpretations between pilots and ATC, while in Air France 447, high-altitude stall responses were interpreted differently depending on assumptions about system reliability and training scripts.

For intelligent system design, these patterns emphasize that open-endedness should be preserved rather than eliminated during annotation and model evaluation. Majority-vote aggregation collapses critical interpretive diversity, especially under moderate-constraint scenarios where cross-role conflicts carry the greatest operational significance. Annotation strategies should deliberately sample across professional roles to capture the spectrum of legitimate interpretations, preserving the tension between traffic management and aircraft control priorities. Similarly, evaluation metrics should reward systems that maintain multiple concurrent intention narratives in ambiguous scenarios, while still converging appropriately when constraints are strong.

Overall, these findings suggest a shift from traditional outcome-driven AI toward interpretive amplification systems. Such systems would enhance professional judgment by surfacing multiple plausible scenarios rather than prematurely resolving ambiguity. By explicitly modeling the 3Cs framework and internal/external scripts, AI can respect the psychological significance of cues and the legitimacy of cross-role differences, supporting more robust situational awareness and safety-critical decision-making.

7 Responsible Research

7.1 Ethical Considerations

This research raises ethical concerns primarily related to how scenario-based models of social intention may influence the development of intelligent systems in aviation. Although the goal of this work is not to predict or judge pilot intentions, but rather to expose interpretive variability, the formalization of data collection scenarios inevitably shapes what kinds of interpretation become visible, recordable, and reusable in future systems.

A central ethical risk lies in the implicit normalization of particular intention narratives during scenario construction. Even when multiple interpretations are presented, the act

of selecting scenarios, defining contextual cues, and framing professional perspectives privilege certain interpretations over others. If such scenarios are later used to train intelligent systems, these systems inherit biases toward institutionally favored perspectives, and hence, contributes to automation bias.

A further ethical tension concerns epistemic authority: who is entitled to define which interpretations of behavior are considered plausible. By categorizing professional perspectives, researchers inevitably make judgments about which viewpoints are sufficiently rational to be included. In aviation, where responsibility is distributed but accountability is often individualized, codifying interpretation patterns into technical artifacts risks oversimplifying the situated expertise of professionals and shifting interpretive authority from human actors to algorithmic systems. This becomes especially critical in a situation where such systems are perceived as objective, despite being grounded in selective assumptions.

7.2 Reproducibility

From a reproducibility perspective, this research adopts a transparent, literature-driven methodology intended to be replicable across domains. The search strategy, query clusters, inclusion and exclusion criteria, and contextual dimensions are explicitly documented, and hence, the process is easily reconstructable. Even though the qualitative synthesis and scenario construction necessarily involve interpretive judgment, all assumptions, and explicit articulation of contextual dimensions have been documented throughout this paper, making it reproducable by anyone showing interest to this topic. By grounding scenario design in documented scripts, and observable cues, the methodology supports methodological rigor.

7.3 Use of AI in the Writing Process

Large Language Models (LLMs) served exclusively as assistance tools throughout the writing process. It was primarily used for text rephrasing and generating tables to present relevant data. The author reviewed and verified all LLM-generated content.

8 Conclusion

This paper examined how contextual variation in aviation can be systematically characterized to support the design of data collection scenarios that enable intelligent systems to capture multiple valid interpretations of social behavior. Addressing a gap in both aviation and AI literature, the study argued that intention recognition in safety-critical environments cannot be reduced to a single ground-truth mapping between observable behavior and inferred intent. Instead, intention interpretation is inherently open-ended, shaped by contextual cues, psychological characteristics, situational classes, and the internal and external scripts held by different observers.

To operationalize this insight, the paper introduced an integrated framework combining the 3Cs model of situational analysis with script theory. This integration provides a structured method for tracing how identical observable cues can be transformed into divergent yet professionally legitimate intention narratives across roles. Through a literature-based synthesis, the framework identified key contextual dimensions that systematically influence interpretation and demonstrated how these dimensions can be manipulated through scenario design. The resulting scenario-first methodology reverses the traditional outcome-driven approach to system evaluation by exposing interpretive ambiguity prior to modeling rather than collapsing it after the fact.

The case studies illustrated how holding observable behaviors constant while varying script configurations reveals unrealized intentions, cross-role conflicts, and training-induced mismatches that aviation professionals routinely manage in practice. These examples show that scenario design is not a neutral preparatory step, but a foundational epistemic process that determines which forms of interpretive diversity are preserved and which are suppressed. From this perspective, aviation intelligent systems should be designed not to eliminate interpretive complexity, but to preserve and surface it in ways that support professional judgment, safety, and coordination.

9 Future Work

This research is exploratory in nature and therefore subject to several limitations that define clear directions for future work. The literature-based methodology, while appropriate for conceptual framework development, constrains empirical coverage and introduces cultural and linguistic biases through its reliance on English-language, predominantly Western aviation sources. In addition, the case studies are derived from secondary literature rather than direct operational observation, limiting ecological validity and professional validation.

Future research should therefore prioritize empirical validation through controlled studies involving aviation professionals, including pilots, air traffic controllers, and maintenance personnel. Such studies would enable systematic testing of the proposed contextual dimensions and assess how different professional groups construct, negotiate, and evaluate intention narratives in real and simulated scenarios. Cross-cultural investigations, supported by international collaboration, would further examine how professional scripts and interpretive norms vary across aviation cultures.

Beyond validation, future work should examine the formal limits of contextual modeling, particularly in degraded operational environments where established scripts may break down and interpretation becomes unstable. Additional research is also needed to explore how scenario diversity and interpretive open-endedness can be preserved without producing unmanageable complexity for annotation, modeling, and system evaluation.

Finally, longitudinal research should investigate how scenario design choices influence downstream research practices, system architectures, and explanation strategies in intelligent systems. Such work would clarify how scenario-first methodologies shape not only data collection, but also the epistemic assumptions embedded in intelligent system development over time.

References

[1] Y. Zeng, Y. Sun, Y. Jie, and X. Liu, "A pilot control intention recognition method based on EEG in simulated flights," *Biomedical Signal Processing and Control*, vol. 112, pp. 1–4, 2026.

[2] N. A. Stanton, R. Stewart, D. Harris, R. J. Houghton, C. Baber, R. McMaster, P. M. Salmon, G. Hoyle, G. Walker, M. S. Young, M. Linsell, R. Dymott, and D. Green, "Distributed situation awareness in dynamic systems: Theoretical development and application of an ergonomics methodology," *Ergonomics*, vol. 49, no. 12-13, pp. 88–101, 2006.

[3] G. L. Clore and J. E. Palmer, "Affective guidance of intelligent agents: How emotion controls cognition," *Cognitive Systems Research*, vol. 10, no. 1, pp. 21–30, 2009.

[4] Anonymous, "Evaluating model explanations without ground truth," arXiv:2505.10399, 2025, preprint. [Online]. Available: https://arxiv.org/abs/2505.10399

[5] H. Artman and C. Garbis, "Situation awareness as distributed cognition," in *Proceedings of the 9th European Conference on Cognitive Ergonomics (ECCE'98)*, Limerick, Ireland, 1998.

[6] W. Irwin and T. Kelly, "Airline pilot situation awareness: presenting a conceptual model for meta-cognition, reflection and education," *Aviation*, vol. 25, pp. 50–64, Aug. 2021.

[7] J. F. Rauthmann and R. A. Sherman, "Conceptualizing and measuring the psychological situation," in *Emerging Approaches to Measuring and Modeling the Person and Situation*, D. Wood, P. Harms, S. Read, and A. Slaughter, Eds. Amsterdam, The Netherlands: Elsevier, 2020.

[8] R. C. Schank and R. P. Abelson, *Scripts, Plans, Goals and Understanding: An Inquiry into Human Knowledge Structures*. Hillsdale, NJ, USA: Lawrence Erlbaum Associates, 1977.

[9] F. Cabitza, A. Campagner, and V. Basile, "Toward a perspectivist turn in ground truthing for predictive computing," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 6, p. 6860–6868, Jun. 2023. [Online]. Available: http://dx.doi.org/10.1609/aaai.v37i6.25840

[10] Air Line Pilots Association, "Aircraft accident report: Tenerife, canary islands, march 27, 1977 — human factors report on the tenerife accident," Air Line Pilots Association, Engineering and Air Safety, Washington, D.C., USA, Accident Report, 1977, pan American World Airways Boeing 747, N736PA & KLM Royal Dutch Airlines Boeing 747, PH-BUF.

[11] X. Peng, Q. Niu, Y. Liang, Y. Luo, N. Lu, and X. Li, "Effects of unexpected event urgency and flight scenario familiarity on pilot trainees' performance and stress responses," *Frontiers in Physiology*, vol. 16, Jul. 2025.

[12] M. D. Babu, D. V. JeevithaShree, G. Prabhakar, K. P. S. Saluja, A. Pashilkar, and P. Biswas, "Estimating pilots' cognitive load from ocular parameters through simulation and in-flight studies," *Journal of Eye Movement Research*, vol. 12, no. 3, Sep. 2019.

[13] R. Anand, "A psychological perspective on aviation disasters through ecological and social lenses," in *Digital Transformation in Aviation Industry Operations*, 1st ed. Routledge, 2025, pp. 245–259.

[14] G. Masi, G. Amprimo, C. Ferraris, and L. Priano, "Stress and workload assessment in aviation: A narrative review," *Sensors (Basel)*, vol. 23, no. 7, pp. 9–13, Mar. 2023.

[15] D. Muñoz-Marrón, "Crew resource management (crm): A historical overview from applied psychology," *Papeles del Psicólogo / Psychologist Papers*, vol. 39, no. 3, pp. 191–199, 2018.

[16] A. C. Garcia, "Air traffic communications in routine and emergency contexts: A case study of flight 1549 'miracle on the hudson'," *Journal of Pragmatics*, vol. 106, pp. 57–71, 2016.

[17] D. Niedermeier, J.-P. Buch, F. Mohrmann, and U. Durak, "Simulating the unexpected: Challenge-centric simulator scenario design for advanced flight crew training," Jan. 2018, pp. 3–12.

[18] P. A. Simpson, "Naturalistic decision making in aviation environments," DSTO Aeronautical and Maritime Research Laboratory, Melbourne, Vic., Australia, General document, DSTO-GD-0279, Jan. 2001, [Online]. Available: http://www.dsto.defence.gov.au/corporate/reports/DSTO-GD-0279.pdf.

[19] K. Parnell, R. Wynne, K. Plant, V. Banks, G. Thomas, and N. Stanton, "Pilot decision-making during a dual engine failure on take-off: Insights from three different decision-making models," *Human Factors and Ergonomics in Manufacturing Service Industries*, vol. 32, pp. 4–13, Nov. 2021.

[20] Bureau d'Enquêtes et d'Analyses pour la sécurité de l'aviation civile, "Final report on the accident on 1st june 2009 to the airbus a330-203 registered f-gzcp operated by air france flight af 447 rio de janeiro–paris," BEA, Paris, France, Final Report, Jul. 2012. [Online]. Available: https://www.faa.gov/sites/faa.gov/files/AirFrance447_BEA.pdf

[21] N. Oliver, T. Calvard, and K. Potočnik, "Sensemaking and control at the limit: The air france 447 disaster," *Academy of Management Proceedings*, vol. 2016, pp. 5–28, Jan. 2016.

[22] B. Kirwan, "Human factors requirements for human–ai teaming in aviation," *Future Transportation*, vol. 5, no. 2, 2025.

[23] J. Korentsides, J. R. Keebler, C. M. Fausett, S. M. Patel, and E. H. Lazzara, "Human–ai teams in aviation: Considerations from human factors and team science," *Journal of Aviation/Aerospace Education & Research*, vol. 33, no. 4, 2024.

[24] X. Mao, L. Ding, X. Sun, L. Pang, Y. Deng, and X. Wang, "Development and implementation of a pilot intent recognition model based on operational sequences," *Aerospace*, vol. 12, pp. 18–19, Aug. 2025.

[25] D.-T. Pham, H. Ali, K. Fennedy, M.-H. Hsieh, S. Alam, and V. N. Duong, "Human–ai hybrid paradigm for collaborative air traffic management systems," in *Proceedings of the SESAR Innovation Days 2024*, Rome, Italy, Nov. 2024, nov. 12–15. [Online]. Available: https://www.sesarju.eu/sites/default/files/documents/sid/2024/papers/SIDs_2024_paper_087%20final.pdf.

[26] B. Matthews, I. Barshi, and J. Feldman, "An approach to identifying aspects of positive pilot behavior within the aviation safety reporting system," in *2023 IEEE/AIAA 42nd Digital Avionics Systems Conference (DASC)*, 2023, pp. 1–6.